International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 6, Issue 3, Jun 2016, 19-38

DESIGN AND REALIZATION OF ON-LINE, REAL TIME WEB USAGE DATA MINING AND RECOMMENDATION SYSTEM USING BAYESIAN CLASSIFICATION METHOD

ADENIYI, D.A, WEI, Z & YONGQUAN, Y

Department of Computer Science & Technology, College of Information Science & Engineering, Ocean University of China, Qingdao, Shandong, China

ABSTRACT

In recent time, there is an increasingly growth in the volume of information available in electronic forms and databases, this therefore makes locating relevant information to be tedious and time consuming. In this paper we have used a unique toolsets to exploit web usage data mining technique to identify a client/visitor's navigation pattern of a particular website specifically, the Really Simple Syndication (RSS) reader's web site, based on the user's current behavior by acting upon the user click stream data, in order to provide tailored information to the individuals so as to ease navigation on the site without too many choices at a time. The Bayesian classification has been trained to be used online and in real time to identify active user click stream data, matching it to a particular user group and recommends a tailored browsing options that satisfies the need of the user at a given period. To achieve this, data mart of user's RSS address URL data extracted from the server database was developed. Experimenting with our work shows that the scalability problem peculiar to this type of system can be overcome through our approach and our results demonstrated that the recommendation system powered by Bayesian classification model can produce accurate, faster and efficient Real-Time recommendation to the client consistently.

***KEYWORDS:** Data Mining, Bayesian Classification, Data Mart, On-Line, Real-Time, Recommendation*

Received: Mar 02, 2016; **Accepted:** Apr 21, 2016; **Published:** Apr 26, 2016; **Paper Id.:** IJCSEITRJUN201603

Original Article

INTRODUCTION

Data mining is the search for valuable information in a large volume of data in order to discover regularity and pattern hidden in data (Niyat *et al*., 2012). Web usage mining can be described as the application of data mining techniques to discover and extract interesting knowledge from a given web site; it is aimed at discovering regularities and patterns in the structure and content of web resources as well as to determine the occurred link-connection on the web site visited by the client. (Niyat *et al*., 2012; Bounch *et al.*, 2001; Xuejuu *et al.*, 2007). Etzioni, was believed to have first came up with the term web mining in his paper titled "The World Wide Web: Quagmire or Gold mine" in 1996, and since then it has caught attention of researchers world over (Resul and Ibrahim, 2008). In recent years, enormous progress has been made in the area of web mining, specifically of web usage mining. Over 400 papers have been published on web mining, since the early papers published in the mid 1990's (Federico and Pier, 2005).

In today's information society, data mining techniques are gaining more popularity for extracting information from databases in different areas, specifically, web log database, this is probably due to its efficiency and capability of working on varieties of databases and amazing results produced at the end of the mining. (Resul

and Ibrahim, 2008; Jiawei and Micheline, 2006; Krzyszlof *et al*.,2007). The newly developed RSS reader's website was meant for reading daily news world-wide, but up till this moment it is not capable of identifying client navigation pattern and cannot offer acceptable real-time response to the web user's needs, this makes finding the appropriate news to become tedious and time consuming. This therefore, makes the benefit of on-line services to become limited. This study is meant to assist the web designer and administration to rearrange the content of the website in order to improve the impressiveness of the web site by providing online real-time recommendations to the client, in order to provide users with the information they want without expecting them to ask for it explicitly.

The study is aimed at designing and developing an online, Real-Time pattern discovery and recommendation system with online pattern matching based on data mart technology. The system will be able to recommend a unique set of objects that satisfy the need of each active user based on the user's current behavior by acting upon the user's click stream data on the newly developed Really Simple Syndication (RSS) reader site, such access and navigation patterns or models are extracted from the historical access data recorded in the User's RSS address URL database, using suitable data mining techniques. The Bayesian classification method was used to investigate the URL information as it relates to the RSS reader web site. (Resul and Ibrahim, 2008). The developed system will be implemented on the newly developed RSS web site. For instance, if a user seems to be searching for sport news on Punch Nigeria Newspaper on his visit to the RSS reader site, more Sports News headlines from other dailies such as China Dailies Sport news will be recommended with the required feed needed to be added to the User's profile in order to access such news headlines. To achieve this, data mart of log data extracted from the Users RSS address database was developed. This is as a result of the fact that the raw users' log database files extracted is not well structured, so it cannot be used directly for data mining. In designing the data mart, the User's RSS address URL database information was consolidated, cleaned, selected and prepared for the data mining analysis. The data acquisition and model extraction operation was carried out using database management software, i.e., the MySQL 2008 (MySQL Corporation, 2008).

The process of the development of the web usage mining and recommendation application was done by adopting the Java programming language with Net Beans as the editor and compiler (Net Beans IDE, 2008). The interpretation and graphical presentation of the result obtained is carried our using the MATLAB Software (Math works incorporation, 1984 – 2011). A thorough presentation of the experimental result was also carried out. The architecture of the overall system is shown in figure 1

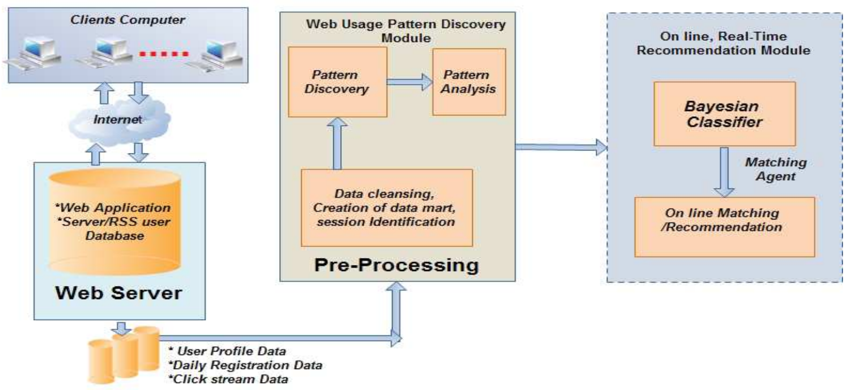

Figure 1: The Overall Web Usage Data Mining System Process Flow

RELATED WORK

This section deals with review of related works pertinent to this study, pointing out similarities and differences with our work. The review is organized into sub sections as follows:

Web Data Mining

Web data mining is a branch of applied artificial intelligence that deals with storage, retrieval and analysis of web log files to discover user accessing patterns of web pages. (Resul and Ibrahim, 2008).

Xuejuu *et al.,* (2007), identified three research areas in web data mining which includes:

- **Web Structure Mining:** This is concerned with investigation of hyperlink structure of the web, so as to discover relevant connectivity patterns usually on the mining of HTML or XML tags.
- **Web Content Mining:** This is about the discovery of pattern contains in the web information which includes; HTML pages, video, images, e-mails audio links etc.
- **Web Usage Mining:** This is the discovery of usage pattern in the web data, in order to understand the user navigation behaviors through his/her click streams at a particular time.

Classification of Data Mining System

Researchers have identified two forms of data mining tasks which are: Predictive and Descriptive (Jiawei and Micheline, 2006; Amartya and Kundan, 2007; David *et al.*, 2001). The predictive mining task makes use of current data in a database by performing inference on them, so as to make prediction of future values of interest while, the descriptive task classifies the data in a database by characterizing the general properties of the data. Descriptive finds pattern describing the data in the database in order to present the interpretation to the user (Jiawei and Micheline, 2006; Amartya and Kundan,2007; David 2012). In order to achieve the goal of either prescription or description tasks, Jiawei and Micheline, (2006), identified different data mining techniques that can be applied which includes: Association rule mining, Clustering, Classification etc which can make use of various data mining methods such as Decision tree, Genetic algorithm, Neural network, Bayesian classifier, Rule based classification, Path analysis, etc.

In web usage data mining task, there are different techniques that can be used, but the main point is how to determine which technique is most suitable for the problem at hand (Federico and Pier, 2005). A comprehensive data mining system can adopt a multiple approach or an integrated technique that combines the advantages of a number of individual approaches (Shu-Hsien *et al.,* 2012).

According to Jiawei and Micheline, (2006), Data mining system can be classified using different criteria such as: Kind of Databases mined, Kind of knowledge mined, kind of technique utilized and according to kind of application adapted.

Classification: Classification is a type of data analysis that can be used to produce models that bring out important data classes. It predicts categorical labels. A classifier is an abstract model that describes a set of predefined classes generated from a collection of labeled data (Luca and Paolo, 2013). There are different techniques for data classification which includes Decision tree classifier, Bayesian Classifier, Rule based classifier, Back propagation, etc. (Jiawei and Micheline, 2006).

In this work the Bayesian classifier, specifically the Naïve Bayesian classification method was applied.

Overview of Some Data Mining Techniques

Below is brief overview of some of the techniques used in Data mining according to different researchers in the field.

Decision Tree: Jiawei and Micheline, (2006), described decision tree as a flowchart-like tree structure in which Test of attributes are represented as internal nodes (non leaf node) and the outcome of the test represented the branch and the leaf node usually called terminal node represents the class label. The Topmost node is referred to as the Root. Amartya and Kundan,(2007), in their work uses classification and regression Tree (CART) in constructing a decision tree, both the gini index(g) and entropy value (e_i) were applied as the splitting indexes. In their experiment they adopted a given set of values, and realized a different sets of results for both the outlook, windy, temp, humidity, and execution time. It The Outlook was discovered to be the best splitting attribute in each cases, with the same order of splitting attributes for both indices (Adeniy et al 2014).

The major problem of the decision Tree algorithm is in the constraint that the training tuples should be located in memory, this therefore makes its construction inefficient in the case of very large data, due to exchange of the training tuples between the main and cache memories. The Bayesian method is capable of overcoming this challenge, since it is more scalable and has the capability to handle training data that are too large to fit in memory (Amartya and Kundan,2007)

The SOM Model: Xuejuu *et al.,(*2007), explore the use of the self organizing map (SOM) or Kohonen neural network model, to model customers navigation behavior. In their work, clusters of queries were created with the model based on user sessions as extracted from web log. Eeach cluster represents a class of users with similar characteristics, so, as to find and recommend the products of interest to a current user on real-time basis. Xuejuu *et al.,(* 2007), further compared the performance of the SOM model with that of K-Means model and discovered that the SOM model outperformed the K-means model with the value of correlation co-efficient of the SOM model scoring twice that of K-means result.

The propose work shares basically the same objective as SOM, but with the difference that its construction is based on building and determining of users' profiles online. This thereby makes real time responses and recommendations faster while in the SOM, the user profiles were pre- determined offline by the offline usage pattern discovery module.

The Path Analysis Model: This is a means of determining the effect of independent factors on dependent factors (Resul and Ibrahim, 2008). Resul and Ibrahim, (2008), in their paper anticipated the use of path analysis technique to examine the URL information of access logs of the Firat University web server's web log files, in order to discover user accessing patterns of the web pages, so as to improve the impressiveness of the web site. They further explain that the application of the said method provides a count of the number of time a link occurred in the data set together with the list of association rules which help in understanding the path that administrators take as they navigate through the Firat University web site.

The Path analysis model uses clients' previous visits' information to determine current user's interest in order to proffer recommendation to the user. The proposed approach shares the same goal of recommendation but with different approach which is based on user's current navigation behaviors rather than previous click behavior as found in the path analysis method.

The K-Nearest Neighbor (KNN) Aalgorithm: (Killian *et al* (ND)), in their work shows how to learn a Mahanalobis distance metric for K-nearest neighbor classification by semi definition programming. The result obtained shows a test error rate of 1.3% on the MNIST handwritten digits. (James *et al.,* 1985), in their own work developed a fuzzy version of the K-NN algorithm by introducing the theory of fuzzy set into K-nearest neighbor technique. The performance comparison between the fuzzy version and the Crisp version result from their experiment shows that, the fuzzy algorithm have low error rate when compared with its counterpart.

Our work shares essentially the same goals as KNN, but differs in its construction, which has to do with the prediction of class membership probabilities rather than focus on local neighborhood as it is in K-NN classification. Performance comparison between our work and K-NN shows that the adoption of Bayesian classification model can bring about a more accurate, faster and efficient recommendation than the K-NN model.

Bayesian Classifier Model: In the work of Rivas *et al.,(*2011), the decision rules, Bayesian networks, support vector machines and classification trees techniques were used to model accident and incidents in two companies, in order to identify the cause of accident. Interview was conducted and data collected were modeled. The result when compared with statistical techniques shows that the Bayesian network and the other method applied is more superior to the statistical techniques. Rivas *et al.,(*2011), explains further that the Bayesian/K2 network is more advantageous as it allows what-if analysis on the data, which make the data to be explore deeply.

In our work, the Bayesian classifier model was adopted, the result of which shows that the Bayesian model especially the naïve Bayesian classifier yields a better result and exhibits a competitive advantage over many other data mining algorithms. It has as well been tested and proved to be capable of overcoming some of the problems with other available algorithms.

Significance of the Study

Available published literature shows that web based recommendation systems are becoming popular; notwithstanding, there are still many problem areas that requires solutions. It has been observed that most existing data mining algorithms are faced with scalability problem and lack some capability when dealing with online, real time search driven web site. Likewise, the recommendation quality and accuracy of some are uncertain. Likewise, some performed poorly when dealing with a very difficult classification task. Some recommendation system can have poor run time performance, if the training set is too large since some have all the work done at run time.

To overcome the problems stated, the following solutions were made through our system.

- Scalability problems common to many existing recommendation system such as the decision tree algorithm were overcome through combine on-line pattern discovery and pattern matching for real time recommendation.
- Our results indicate that the adoption of the Naïve Bayesian model can bring about a more accurate recommendation that outperformed many other existing models. In most cases the precision rate or quality of recommendation by our system is equal to or better than 85%, meaning that over 85% of news recommended to a user will be in line with her immediate browsing interest, making support for the browsing process more genuine instead of a simple reminder of what the user was interested in on her previous visit to the site or from her recorded user profile as found in path analysis technique.

- The proposed recommendation system collects the active users' click stream data, matches it to a particular user's group and then generate a set of recommendations to the client based on her current interest at a faster rate. This therefore alleviate the computational complexity problem or of bottleneck caused by system computing load when handling scaled web sites at a peak visiting time.
- The proposed system provides a precise recommendation to the client based on her current click stream data, thereby reduces time spent in finding the right news or information which is usually as a result of presentation of many irrelevant choices to the user at a time as it is in many existing systems that uses navigation history or recorded user profiles

The naïve Bayes model is more efficient than most other algorithm such as K-NN, since the probabilities may be computed at learning time and use it in the actual classification, therefore overcoming the problem of poor runtime performance.

Finally, the Bayesian model is capable of handling a very difficult classification task unlike the K-NN model.

Hence, the proposed approach addresses the issues and provides a useful, accurate, faster and efficient web usage classification and recommendation model.

METHODOLOGY

This section describes in detail the realization of the web usage data mining system. The applications of the proposed methodology for analyzing the Users RSS address URL database of the RSS reader site were presented. An online, Real-time recommendation expert system has been developed to assist the web designer and administrator to improve their web site by recommending a unique set of objects that satisfy the need of an active user based on his/her current click stream.

Overview of Steps in Performing Web Usage Data Mining Tasks

According to (Amartya and Kundan, 2007; David *et al.,* 2001), data mining task can be categorized into different stages relating to the objective of the individual who is analyzing the data. The objective of our system is to design and develop a Real-Time recommendation system with on-line pattern matching. The system is aimed at recommending a unique set of objects that satisfy the need of each active user based on the user's current behavior by acting upon the user's click stream data on the RSS site, so as to avoid irrelevant recommendations common in most of the existing recommendation systems that are based on user's previous visit to the site.

The overview of the task for each steps are presented in detail in four sub sections as follows:

Data Acquisition, Pre-Processing and Data Mart

Data Acquisition: The first task in web mining application is data acquisition. Data can be collected from three main sources for the purpose of Web usage mining, this includes: (i). Web server (ii). Proxy server and (iii). Web Client (Federico and Pier, 2005; Dario et al., 2013). For the purpose of this study the Web server source was chosen as it is the richest and most common source of data, moreover, it can be used to collect large amount of information from the log files and databases that they represent, which usually contain basic information such as name and IP address of the remote host, date and time of request, the click streams, etc., all which are represented in standard format such as common Log format,

Extended Log format and Log ML format (Federico and Pier, 2005). The access and navigation patterns or models are extracted from the historical access data recorded in the Users address URL database of the RSS reader site.

The data is so large, as it contain so many detailed information such as date, time at which activities occur, server name, Ip address, user name, password, dailies name, required fees, news headlines and contents, etc., as recorded in the database, in fact the original document is about 5,285 pages.

Data Pre-Processing: The raw Users RSS address URL Database extracted is made up of text file that contain a large amount of information concerning queries made to the web server which contains irrelevant, incomplete and misleading information for mining purposes (Xuejuu *et al.,* 2007).

According to Resul and Ibrahim, (2008), data preprocessing is the cleansing, formatting and grouping of raw web log files into meaningful session for the purpose of utilizing it for web usage analysis.

Data Cleansing: Data cleansing is the process of eliminating irrelevant/noisy entries from the access database (Michal, Jozef and Peter, 2012). The data cleansing operations carried out on extracted database includes:

- Removal of entries that have “error” or “failure” status
- Identification and removal of some access log data that are generated by automated programs from the access log file and proxies.
- Removal of request for pictures files associated with request for a page and request include Java script(.JS), style sheet files etc.
- Removal of entries with unsuccessful HTTP status codes, etc.

Data Mart: The development of a data mart of log data is a requirement for data mining operation as the raw log file is not a good starting point for data mining. Therefore a separate Data mart of User’s RSS address URL database was developed using relational Database management software, MySQL (MySQL Corporation, 2008). According to Two Crown Corporation, (1999), the data mart may be a logical subset of a data warehouse, if the data warehouse DBMS can support more resources that will be required of the data mining operation, otherwise a separate data mining database will be required.

Transaction Identification

Xuejuu *et al.,* (2007) described transaction identification as the process of creating meaningful clusters of references for each identified users, it represents a user’s navigation behavior as a series of click operations by the particular user in a time succession; this is usually referred to as click stream. The click stream can further be divided into units of click descriptions called session.

Session Identification: According to Xuejuu *et al.,* (2007), a session (also referred to as a visit) is a collection of user click to a single web server. At the end of the data cleansing operation the log entries are partitioned into sessions (Michal, Jozef and Peter, 2012), to do this, Xuejuu *et al.,* (2007), suggested the use of cookies for the purpose of identifying individual users, so as to get a series of clicks within a time interval for an identified user. Two clicks can be included in one session, if the time interval between them is less than a specified period. In Cooley *et al.,* (1999) model, each user session can either be a single transaction made up of many page references or a set of many single page reference

transactions. The time-out between every two clicks of a user as calculated by Catledge and Pitkow, (1995) indicate the mean value to be 9.3 mins and by adding 1.5 derivations to the mean, we now have a maximum time-out of 25.5 mins for two adjacent clicks in one session. While, Cooley *et al.*,(2000) also, adopted 30 mins as maximum time-out for the same purpose in their work.

Pattern Discovery

At the completion of data pre-processing and transaction/session identification as described in section 3.1.1 and 3.1.2, the next thing is to group the users based on similarities in their profile and their search behavior. There are series of web usage mining techniques that can be used for pattern discovery and recommendations such as path analysis, clustering, associate rules, etc. In our work we have experimented with Bayesian classification technique, specifically, the Naïve Bayesian classification technique as described in section 3.2 so as to observe the pattern of user behavior and click stream from the pre-process to web log stage.

Pattern Analysis

Pattern analysis is aimed at removing irrelevant rules or statistics in order to extract interest rules or statistics from the result of the pattern discovery process. This stage provides the tools for the conversion of information into knowledge. The result of our work is stored in a data mart developed and implemented using the MySQL DBMS software specifically created for the purpose of web usage data mining. The data mart is populated from raw user's RSS address URL database of the RSS reader's site that contains some basic fields needed. The results of our experiment are presented in section 4.

Our Approach

The Bayesian Classification has been chosen, because from our work, it has shown that it has a minimum error rate in comparison to all other classifier. The Bayesian classifier also provide a theoretical justification for other classifiers, study comparing classification algorithms shows that the Naïve Bayesian classifier can be comparable in performance with other classifiers such as decision tree, selected neural networks, more so, Bayesian classifier has high degree of accuracy and speed when used in a large databases (Jiawei and Micheline, 2006).

Bayesian Classification technique

Bayesian classification technique is a statistical classifier technique that can be used to predict membership probability, that is, the probability that a given tuple belongs to a specific class (Jiawei and Micheline, 2006).

The Bayesian classification is derived from the Bayes' Theorem. The Bayes' theorem is the hand work of Thomas Bayes, an English clergyman who in the 18^{th} century did early work on probability and decision theory. This was later named after him, hence the name Bayes' theorem. According to Jiawei and Micheline, (2006). The Bayes' classifier comes in two forms viz:

(a). Naïve Bayesian classifier and (b) Bayesian Believe networks

The Naïve Bayesian classifier assumes that the effect of an attribute value on a given class is independent of the value of the other attributes. This is referred to as "class conditional independence" while Bayesian Belief network are graphical models that allows representation of independencies among subset of attributes.

Basic Probability Notation of Bayes' Theory Using Naïve Bayesian Classification

Bayes' theory assumes X to be a data tuple, X is referred to as "evidence", usually described by measurements made on a set of n attributes.

Let C be a specific class

Let H be some form of hypothesis, such that, the data tuple X belongs to a specific class C.

We want to determine the probability that tuple X belong to Class C, giving the attribute of X i.e. P(H|X) for classification problems. P(H|X) is referred to posterior probability or a posteriori probability of H, conditioned on X (Jiawei and Micheline, 2006).

In our experiment, assuming that the given data tuples are limited to a client depicted by the attribute Daily Name, Daily Type and News category, we also assume that X is a client with Dele as his username and DL234 as password.

Assuming that H is the hypothesis that our client will read similar news category as his selected news category during his current browsing session, Then P(H|X) shows the probability that client X will read similar news category given that we know his daily's name and news category.

On the contrary, the Prior probability, P(H),or priori probability of H in our work is the probability that any client will like to read any news category irrespective of his selected Dailies name and news category or any other information as the case may be.

More so, P(X) is the prior probability of X. In our experiment, it is the probability that a visitor from our set of clients clicked a particular daily name and news category. We may estimate P(X), P(X|H) and P(X) from the data provided as shown below using Bayes' theorem as it can provide a way of calculating the posterior probability, P(H|X), from P(H), P(X|H) and P(X).

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad \ldots\ldots\ldots \qquad \text{equation 3.1}$$

This is called the Bayesian theory (Jiawei and Micheline, (2006).

The Working of Naïve Bayesian Classifier

Explanation on the working of naïve Bayesian classification is narrated below:

- Given that D is a training set of tuples and their related class labels, representing each tuple by an n-dimensional attribute vector such that X = (x_1, x_2,………x_n) where n is the number of measurements made on the tuple, with A_1,A_2, ………,A_n represent the attributes.
- Given a tuple X with m number of Classes, c_1,c_2,……c_m, The Naïve Bayesian classifier will predict that X is a member of a class with highest posterior probability. Ie X belong to class Ci if and only if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m$, $j \neq i$

From equation 3.1 (Bayes' theorem), we can maximize $P(C_i|X)$, where the class C_i is called maximum posteriori hypothesis as follows:

$$P(Ci|X) = \frac{P(X|Ci)\,P(Ci)}{P(X)} \ldots \qquad \text{equation 3.2}$$

Note: Only $P(X|C_i)$ $P(C_i)$ needs to be maximized so far P(X) is constant for all classes.

- There is need to apply the naïve class conditional independence assumption when encountering with set a of data with many attributes in order to reduce too much computation of $P(X|C_i)$. To this effect we presume that the attributes values are conditionally independent of one another. So, we have

$P(X|C_i) = \prod_{k=1}^{n} P(Xk|Ci) = P(X_1|C_i) \text{ X } P(X_2C_i) \text{ X } \ldots\ldots \text{X } P(X_n|C_i)\ldots$ equation 3.3

This makes it easy to estimate the probabilities of

$P(X_1|C_i)$, $P(X_2C_i)$ …………. $P(X_n|C_i)$ from the given training tuples.

Note: that X_k is the value of attribute A_k for tuple X.

Now, we have to determine whether the attributes is categorical or continuous valued. To compute P(X|Ci), we determine that: If $P(X_k|C_i)$ is the number of tuples of class C_i in D with value X_k for A_k, divided by $|C_i,D|$, the number of tuples in class C_i in D, then A_k is said to be categorical; otherwise A_k is continuous–valued.

A continuous-valued attributes usually assumed a Gussian distribution with mean μ and standard deviation б, defined as follows:

$$g(x, \mu, б) = \frac{1^{e^{-\frac{(x-\mu)^2}{2б^2}}}}{\sqrt{2Лб}} \ldots\ldots\ldots \qquad \text{Equation 3.4}$$

The mean μC_i and standard deviation $бC_i$ of the values of attribute A_k for training tuples of class C_i must be computed and substituted into equation 3.4 together with x_k, so as to estimate $P(x_k|C_i)$.

This gives us $P(X_k|C_i) = g(X_k,\mu C_i,бC_i)$ equation 3.5

- To predict the class label of X, $P(X|C_i)P(C_i)$ must be evaluated for each class Ci.

The naïve Bayesian classifier predicts that tuple X belong to class Ci if and onlt if

$P(X|C_i)P(C_i)> P(X|C_j)P(C_j)$ for1 $\leq j \leq$ m, $j \neq I$ ………….. equation 3.6

This is to say that the class label with class Ci for which $P(X|C_i)P(C_i)$ is the maximum is predicted. The algorithm for the Bayesian classifier model is shown in figure 2.

```
(1) Recommendation-set ={ }; // set of
    recommendation is empty
(2)  C= { },such that C= {C1, C2, .................,
    Cm}; // C1 to Cm
              represent different classes
(3)  D={X}; X = (x1 ,x2 .........xn);// Data set class
    label  tuple
(4)       For each client X in the database do
(5)         Repeat
(6) Recommendation-set = X ∈ Ci, iff,
          P(X|Ci)P(Ci)> P(X|Cj)P(Cj) for 1 =j= m, j?i
(7) Return recommendation-set
(8) Untill terminating condition
(9) End for
```

Figure 2: The Algorithm for the Bayesian Classifier Model

Application of Naïve Bayesian Classification Technique to Predict a Class Label in the RSS Reader's Site

Example 1: Given the training data tuple as in table 1, based on user's click streams in the RSS reader's site and the class label suggesting possible recommendations by the developed system.

Table 1: The RSS Readers' Data Mart Class Label Training Tuple

S/no	Daily Name	News Category	Added Required Sport feed	Read/Recommend Related Sport news heading
1	China Daily	Sports	Yes	Yes
2	CNN News	Sports	Yes	Yes
3	China Daily	Politics	No	No
4	Thisday News	Sports	Yes	Yes
5	Punch Nigeria	Sports	Yes	Yes
6	Vanguard News	Entertainment	No	No
7	Vanguard News	Sports	Yes	Yes
8	CNN News	World	No	No
9	New Nigeria	Sports	No	No
10	completeFootbal	Sport	Yes	Yes
11	Punch Nigeria	Politics	No	No
12	China daily	Business	Yes	No
13	Punch Nigeria	Sport	Yes	Yes

The data tuples above are described by the attribute Daily Name, News Category, Add required sport feed and the class label attribute, Read/Recommend related SportNews Headlines has two values {Yes, No}, taking C_1 to be the class, Read/Recommend related SportNews Headlines = Yes and

C_2 to be the class, Read/Recommend related SportNews Headlines = No

So, we have the following tuples to classify:

X = (Daily Name = China daily, News Category = Sports, Addrequired Sport feed = Yes)

To maximize $P(X|C_i)P(C_i)$, for i = 1,2, $P(C_i)$, we can compute the prior probability of each class based on the training tuples as follows:

P(Read/Recommend Related Sports News Headlines = Yes) = 7/13 => 0.538

P(Read/Recommend Related Sports News Headlines = No)= 6/13 => 0.462

We can now compute the following conditional probabilities in order to get the value of $P(X|C_i)$, for i = 1,2,

P(Daily Name = Chinadaily| Read/Recommend Related Sports News Headlines = Yes) = 1/7 => 0.123

P(Daily Name = Chinadaily| Read/Recommend Related Sports News Headlines = No) = 2/6 => 0.333

P(News Category = Sports| Read/Recommend Related Sports News Headlines = Yes) = 7/7 => 1.000

P(News Category = Sports| Read/Recommend Related Sports News Headlines = No) = 1/6 => 0.167

P(Added required Sport feed = Yes| Read/Recommend Related Sports News Headlines = Yes) = 7/7 => 1.000

P(Added required Sport feed = Yes| Read/Recommend Related Sports News Headlines = No) = 1/6 => 0.167

Applying the possibilities above, we have:

P(X| Read/Recommend Related Sports News Headlines = Yes)= P(Daily Name = Chinadaily| Read/Recommend Related Sports News Headlines = Yes) X P(News Category = Sports| Read/Recommend Related Sports News Headlines = Yes) X P(Added required Sport feed = Yes| Read/Recommend Related Sports News Headlines = Yes)

= 0.123 X 1.000 X 1.000; =0.123, Likewise,

P(X|Read/Recommend Related Sports News Headlines = No)= P(Daily Name = Chinadaily| Read/Recommend Related Sports News Headlines = No) X P(News Category = Sports| Read/Recommend Related Sports News Headlines = No) X P(Added required Sport feed = Yes| Read/Recommend Related Sports News Headlines = No)

= 0.333 X 0.167 X 0.167; = 0.009

Since,the Read/Recommend Related Sports News Headlines = Yes, is the maximum, the naïve Bayesian Classifier predicts, Read/Recommend Related Sports News Headlines = Yes for tuple X. Therefore, more sport news can be recommended for the user with tuple X.

Overcoming the Challenges of Zero Probability

There is possibility of encountering zero probability value. For instance in our example there are two classes, viz: Read/Recommend Related Sports News Headlines = Yes and Read/Recommend Related Sports News Headlines = No, In a situation where there is no training tuple for a paticular class, then we may end up with probability with value zero, applying zero in equation 3. 4 will return a zero probability of $P(X|C_i)$, and this will eventually cancel the effect of all other posterior probabilities on C_i involved in the experiment. To overcome this challenge, Jiawei and Micheline, (2006), suggested the application of Laplacian correction or Laplace estimator. The Laplace estimator was named after Pierre Laplace, a french mathematician (1749 to 1827) (Jiawei and Micheline, 2006). In this technique, we simply add 1 to each count, by assuming that our training database D is so large, that adding the 1 will only make an insignificant difference in the estimated probability value hence avoid the case of zero probability value. If we add 1 to each count, say K then K must be added to the correspondence denominator used in the probability calculation (Jiawei and Micheline, 2006).

SYSTEM EVALUATION AND RESULT ANALYSIS

This section applied the result of the experiment conducted to evaluate our system, present and analyse the result so as to evaluate the quality of our recommendation system based on Bayesian classification model. In the previous section we established that a class with highest or maximum priority probability will be predicted and recommendation will be made based on this, for the user with tuple X.

Figure 3, shows sample interface from the on-line recommendation system developed for the purpose with Java Net Beans programming language indicating the active user's click stream, a dialog box presenting his requested headlines and a message box presenting online recommendation based on his current request.

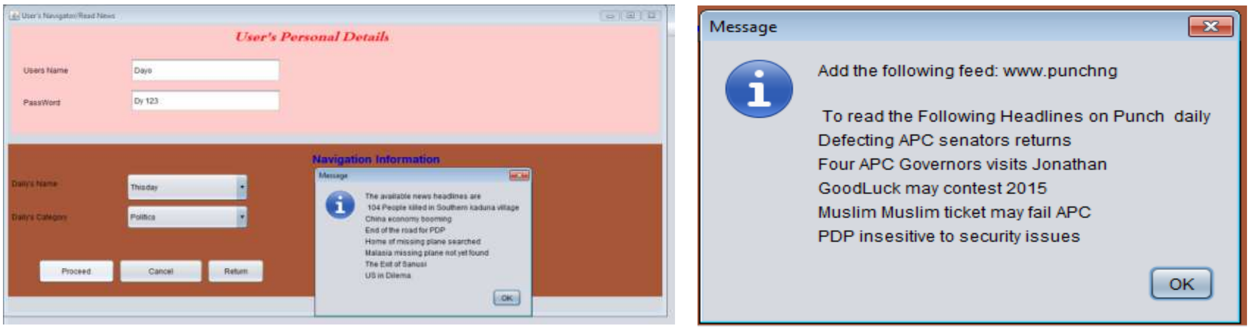

Figure 3: Sample Interface from the On-Line Reccomendation System.The Source code in Java Net Beans Programming Language for the Developed Application is Available on Request

In this study, the number of Class C of user X to be recommended by the recommendation model is set at 8, 8 represent different news categories headlines that could be presented to the active user, based on information from his click stream.

In this work however, the computation of conditional probabilities that produced the values of P(X|Ci) for i = 1,2, , n, was not repeatedly shown, because of size, since the computation follows the same procedures, hence table 2, shows the result of the computations.

A ccording to Godswill, (2006), in real-life analysis, a model performance quality can only be measured by ability to predict correctly the new data set rather than trained data sets in which the model was trained. Godswill, (2006), further stated, that the predictive ability of a model is questionable and can not be used for prediction, even if the model performs well in the training set but performs unsatisfactorily in the test validation data set or new data set.

Presentation of results

The whole process in our example 1. can be repeated for all the available Dailies name, News categories and Added required feeds in order to arrive at recommendations for the user based on his or her current click streams. These calculations produced a stream of data as shown in table 2

Table 2: Recommendation for the user Based on His or her Current Click Streams

S/no	*Class(X)*	*Yes*	*No*
1	P(X) Read/Recommend Related Sports News Headlines	0.123	0.009
2	P(X) Read/Recommend Related Politics News Headlines	0.12	0.068
3	P(X) Read/Recommend Related Business News Headlines	0.14	0.084
4	P(X) Read/Recommend Related Entertainment News Headlines	0.135	0.088
5	P(X) Read/Recommend Related World News Headlines	0.133	0.075
6	P(X) Read/Recommend Related Health News Headlines	0.11	0.1
7	P(X) Read/Recommend Related Technology News Headlines	0.122	0.115
8	P(X) Read/Recommend Related Science News Headlines	0.121	0.089

Analysis of the Result

The MATLAB code (Math works incorporation, 1984 – 2011; Ogbonaya, 2008), used for graphical analysis of the experimental result as in figures 4 to 7 is available on request.

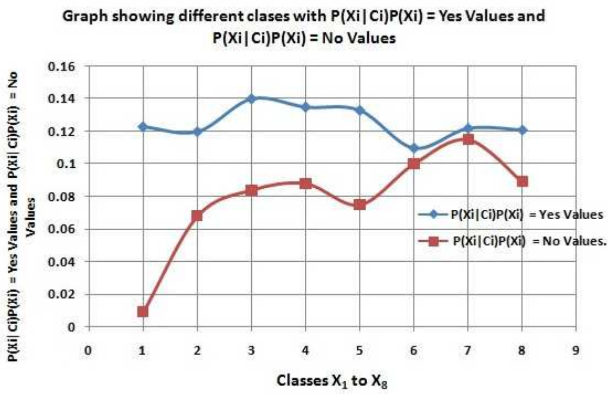

Figure 4: Graph showing the result of class with $P(X_i|C_i)P(X_i)$ = Yes Values and $P(X_i|C_i)P(X_i)$ = No Values

In order to model the visitors click stream in the RSS reader's web site. The Naive Bayesian classification technique of data mining was applied on the extracted web Users RSS address URL Database. The data sets were produced by computing the conditional probabilities of P(Xi|Ci), for i= 1,2, ……, n as shown in example 1. and the data set presented in table 1.

The naïve Bayesian classifier predicts the class label with class Ci for which P(Xi|Ci) P(Ci) is the maximum. Figure 4, clearly shows that the class label with class Ci for which $P(X_i|C_i)$ $P(C_i)$ = Yes, is predicted for users X_1 to X_8. The result shows that while$P(X_i|C_i)$ $P(C_i)$ = Yes has values 0.123, the $P(X_i|C_i)$ $P(C_i)$ = No, has a value 0.009 Therefore, $P(X_i|C_i)$ $P(C_i)$ = Yes is recommended, being the maximum. All through, it is discovered that $P(X_i|C_i)$ $P(C_i)$ = Yes, usually have the maximum value for each user, X_1 to x_8, so it is always recommended. Hence, if $P(X_i|C_i)$ $P(C_i)$ = No, has the maximum values, that means no recommendation will be made to the particular visitor.

Model Comparison

This section briefly compares the performance of the Bayesial algorithm with that of traditionam Euclidean distance K-NN classification algorithms in order to determine the effectiveness of the Bayesian algorithm on our data sets.

Baseline Algorithm

In this section, we briefly describes the K-Nearest Neighbour(K-NN) classifier. Algorithm use as baseline for comparison against the Bayesian classifier algorithm used on our data sets.

The K-Nearest-Neighbor Technique

Jiawei and Micheline, (2006), describes the K-Nearest Neighbor classification algorithm as a straightforward and popular pattern recognition algorithm. It has been used by Amazon.com to provide recommendation to on-line news readers. It learn by analogy i.e., simply by comparing a specific test set with a given training sets that are similar to it. It classifies a given data tuple based on the class of their closest neighbors (Adeniyi, Wei and Yongquan, 2014). Most of the time, more than one neighbor is taken into account hence, the name K-Nearest Neighbor (K-NN), the "K" is usually the number of neighbors taken into account in determining the class a given test tuple belongs (Jiawei and Micheline, 2006).

The K-NN mostly applies either the Euclidean distance or the cosine similarity between the trainning set and the test sets(Adeniyi, Wei and Yongquan, 2014). The Euclidean distance between a training attribute and a test attribute is given as

$$\text{Dist}(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$$

For instance the distance between attribute x_1 and x_2 such that x_1= (3, 5) and x_2= (1, 2),can be derived as

$$= \sqrt{(3-1)^2 + (5-2)^2} \quad , = 3.61.$$

The above expression is used for numeric attributes, for non-numeric attributes, such as name of object, color of object etc. The distance is computed simply by comparing the value of corresponding attributes x_1 with that of x_2. If the values are the same, the difference is taken to be zero (0) otherwise it is taken to beone (1). The next step is to pre-sort(arrange) the distance between the test tuples and the trainning tuples in ascending order of their closenes to the test tuple. The one on the top of the sorted list is selected first and is class will be used to make recommendation to the unknown (test) tuple. See Adeniyi et al, for a more detailed description of Euclidean distance K-NN.

DISCUSSIONS

In this section, the performance of the Naïve Bayesian classifier is compared with that of K-Nearest Neighbour(K-NN) classifier. We experimented with over 100 different attributes and finds out that the K-NN performs well with fewer number of attributes, adding additional attributes always lowers the accuracy unless all the attributes added are all relevant. For instance, the K-NN algorithm with Eucleadan distance reaches its maximum at three attributes and then goes down quickly with addition of more attributes, as can be seen in Table 3and figure 5, therefore leading to a limited number of useful predictions. In fact, according to James et al (1985), in the infinite sample situation, the error rate for the 1-NN rule is bounded above by no more than twice the optimal Bayes error rate and the error rate approaches the optimal rate asymptotically, as the value of K increases.

In the case of Naïve Bayes classifier, we used an experimental settings similar to that used earlier for K-NN. The result shows that for all sample attributes between 3 to 100, the Naïve Bayes classifier has 100% accuracy and after that maintains about 85% accuracy with all attributes as shown in Table 3 and figure 6. So, the clear winner in this experiment is the Naïve Bayes classifier.

In summary, the Naïve Bayes algorithm is suitable for domain with large vocabularies such as the web. The Naïve Bayes algorithm works well in practice, without the need to select a small subset of relevant attributes, despite the fact that it is based on an independent assumption which is almost never present in real data.

The Naïve Bayes classifier is also more efficient than the K-NN, since the probabilities may be computed at learning time and use the results in the actual classification, thereby making it straightforward as well. (Zdravko and Daniel, 2007). In K-NN algorithm, all the work is done at run-time, therefore results in poor runtime performance, if the trainning set is large. This problem is overcome by Bayesian algorithm.

More so, K-NN is very sensitive to irrelevant or redundant features because all features contributes to the similarities and the classification, which if the features are not carefully selected, it might result in inaccurate recommendations. This is usually not the case in Bayesian model.

Finally, the Bayes clasifier can outperforms the K-NN when it comes to a very difficult classification tasks.

Table 3: Comparison of Accuracy of Naïve Bayes and 1-NN with Different Number of Attributes

Number of Attribute	Naïve Bayes Accuracy (%)	1-NN Accuracy (%)
1	85%	70%
2	95%	80%
3	100%	95%
5	100%	90%
10	100%	87%
20	100%	85%
30	100%	85%
50	100%	70%
100	100%	65%
120	85%	50%
130	85%	45%
150	75%	45%
160	75%	45%
165	70%	45%
168	85%	40%

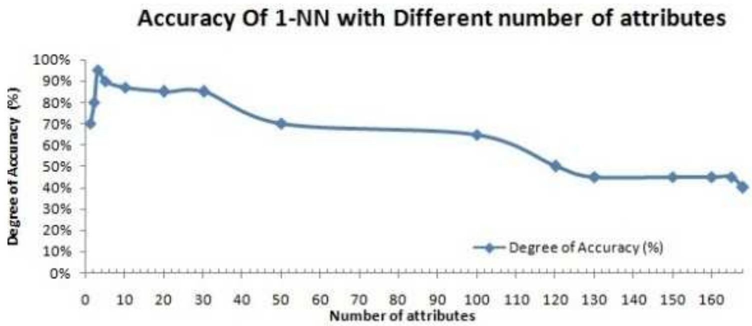

Figure 5: Accuracy of 1-NN with different Number of Attributes

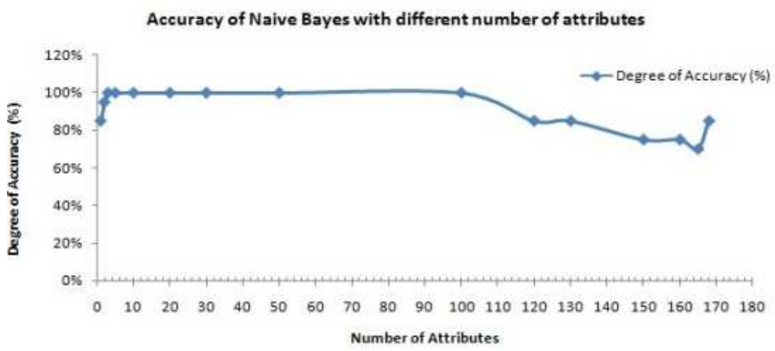

Figure 6: Accuracy Of Naïve Bayes with different Number of Attributes

Summary of Findings

The quality of web a site can be determined by many factors which includes: content, presentation, ease of usage, ease of locating required information, user waiting time, etc. This study demonstrated that web usage data mining can be used to extract knowledge required for providing a Real-Time recommendation service on the web. This is aimed at helping web designers and administrators to improve their web site specially, that of RSS reader site, by observing user behaviour during their visit to the site through their click streams and provide appropriate recommendations to the

individuals in order to ease navigation on the site without too many choices being offered at a time, as well as meeting their needed information without expecting them to ask for it explicitly.

To achive this, raw Users' RSS address URL database of the RSS reader site were extracted, pre-processesd and the Bayesian classification technique was used to investigate the URL information of the Users' RSS address URL database of the RSS reader site, as stored in the data mart created. The results were presented and analysed. The findings of the experimental study can now be used by the designers and the administrators of the web site to plan the upgrade and improvement of the web site.

CONCLUSIONS

This work presents a foundation for online, real-time news recommendation system. The system only needs to collect active user's click stream and match this with a similar class in the data mart, in orders to generate a set of recommendations to the client in a Real-Time basis. The result of our experiment shows that an online, Real-Time recommendations engine powered by Bayesian classification model is capable of producing useful and accurate news recommendation to the client at any time based on her current browsing requirement, rather than information based on his previous visit to the site or predefined user's profile.

Many proposed approaches to creating web-based recommendation systems understudied lack scalability and capability when dealing with search-driven web sites in a real time, online basis. Our approach seeks to overcome some of these problems.

Our recommendation system is capable of generating a set of recommendation to the client at a faster rate through online determination of user profile which make real time response possible.

Our system has also been proved to have capability to overcome the problem of poor run time performance common to many existing models when dealing with large training sets due to its capability to compute the probability at learning time, then use it in the actual classification at run time.

Performance comparison between our system and the K-NN model as shown in Table 3, figure 5 and figure 6, indicates that the adoption of Naïve Bayesian classifier can lead to a more accurate recommendation that outperformed the K-NN model. In most cases the precision rate or quality of recommendation is equal to or better than 85%, meaning that over 85% of news recommended to the client will be in line with his immediate requirements, making support for the browsing process more genuine rather than a simple reminder of what the user was interested in on her previous visit to the site, as it is in Path analysis technique. Therefore, our system is capable of providing a useful, accurate, faster and efficient web usage classification and recommendation model, online and in real-time basis.

RECOMMENDATION FOR FUTURE WORK

We are of the opinion that this study could be taken much further by investigating the users' RSS address URL database of the RSS reader site on a continous basis. In addition, there is the need to research on other data mining techniques, comparing the result with this model so as to determine which one will be more efficient in handling a problem of this nature in the nearest future.

REFERENCES

1. *Niyat, A., Amit, K., Harsh, K., Veishai, A., (2012). Analysis the effect of data mining techniques on database. Journal of advances in Engineering & software, 47(2012) 164-169.[doi:10.1016/j.advengsoft.2011.12.013].*
2. *Bounch, F., Giannotti, F., Gozzi, C., Manco, G., Nanni, M., Pedreschi, D., Renso, C., Ruggier, S., (2001). Web log data warehousing and mining for intelligent web caching. Journal of Data and Knowledge engineering **36**(2001); 165-189. [PH:S0169-023x(01)00038-6]*
3. *Xuejuu, Z., John, E., Jenny, H., (2007). Personalised online sales using web usage data mining. Journal of computer in industry. 58(2007）772-782.[doi:10.1016/j.compind.2007.02.004].*
4. *Resul, D., Ibrahim, T.,(2008). Creating meaningful data from web log for improving the impressiveness of a web site by using path analysis method. Journal of expert system with applications 36(2008) 6635-6644. [doi:10.1016/j.eswa.2008.08.067].*
5. *Federico, M.F., Pier, L.L., (2005), Mining interesting knowledge from weblog: a survey. Journal of Data and Knowledge engineering 53(2005): 225-241.[doi:10.1016/j.datak.2004.08.001].*
6. *Jiawei, H., Micheline, K.,(2006). Data mining concept and Techniques. 2nd edition. Morgan Kaufmann Publishers, Elsevier inc., USA San Francisco, CA 94111,P.285-350.*
7. *Krzyszlof, J.C., Witold, P., Roman, W.S., Lukasz, A.K., (2007). Data mining : A Knowledge discovery approach, Springer Science + Business media, LLC, USA, New York, NY 10013.*
8. *MySQL corporation.,(2008). MySQL Database Management System software. USA MySQL/Oracle corporation.*
9. *Net Beans IDE 7.3., (2008). Net Beans java compiler. USA, Java/Oracle corporation.*
10. *Math works incorporation., (1984 – 2011). MATLAB R2011b(7.13.0.564), License number: 161052, USA, Math works incorporation.*
11. *Amartya, S. and Kundan, K.D., (2007). Application of Data mining Techniques in Bioinformatics. B.Tech Computer Science Engineering thesis, National Institute of Technology, (Deemed University), Rourkela.*
12. *David, H,, Heikki, M., Padhraic, S.,(2001). Principles of data mining. The MIT press, Cambridge. Massachusetts, London, England, p. 2-20.*
13. *Shu-Hsien, L., Pei-Hui, C., Pei-Yuan, H., (2012). Data mining techniques and applications- A decade review from 2000 to 2011. Journal of expert system with applications 39(2012) 11303-11311.[doi:10.1016/j.eswa.2012.02.063].*
14. *Rivas, T., Paz,M., Martins,J.E., Matias, J.M., Gracia,J.F., Taboadas, J.,(2011. Explaining and predicting workplace accidents using data-mining Techniques. Journal of Reliable Engineering and System safety 96(7) 739-747. [doi:10.1016/j.ress.2011.03.006].*
15. *Two crown corporation,(1999). Introduction to Data mining and Knowledge discovery, Third edition. Two crown corporation, 10500 falls road, Potamac, MD 20854, USA. P. 5-40.*
16. *Cooley, R., Mobasher, B., Srivastava J. (1999). Data preparation for mining World Wide Web browsing patterns. Journal of knowledge and information system 1(1). 1-27.*
17. *Catledge, L. D., Pitkow, J., (1995). Characterizing browsing strategies in the world wide web, Journal of computer Networks and ISDN system. 27(6), 1065-1073.[doi: 101016/0169-7552(95)00043-7].*
18. *Cooley, R., Tan, P.N., Srivastava, J., (2000). Discovery of interesting usage patterns from web data. International workshop on web usage analysis and user profiling, ISBN 3-540-67818-2, P.163-182.*

19. *Godswill, C.N., (2006). A comprehensive Analysis of predictive data mining techniques. M.Sc. Thesis, The University of Tennessee, Knoxville.*

20. *Ogbonaya, I.O., (2008). Introduction to Mat lab/Simulink, for engineers and scientist, 2nd edition. John Jacob's Classic Publishers Ltd, Enugu, Nigeria.*

21. *Dario, A., Eleno, B., Giulia, B., Tania, C., Silvia, C., Naeem, M., 2013. Analysis of diabetic patients through their examination history. Journal of expert Systems with Applications. 40(2013)4672-4678). [doi:dx.doi.org/10.1016/j.eswa.2013.02.006].*

22. *David, F.N., 2012. Data mining of social networks represented as graphs. Journal of computer science review. 7(2013)1-34. [doi:10.1016/j.cosrev.2012.12.001].*

23. *Habin, L., Vlado, K., 2006. Combining mining of web server logs and web content for classifying users' navigation pattern and predicting users' future request. Journal of data and Knowledge engineering. 61(2007) 304-330.[doi:10.1016/j.datak.2006.06.001]*

24. *James, M.K., Michael, R.G., James, A.G., 1985. A Fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on system Man and cybernetics, vol. SMC-15 No4.[0018-9472/85/0700-0580$01.00].*

25. *Killian, Q.W., John, B., Lawrence, K.S.,(ND), Distance metric learning for large margin Nearest Neighbor classification. Technical report, University of Pennsylvania, Lavine Hall,3330 Walnut street, Philadelphia, PA 19104.*

26. *Leif, E.P., 2009. K-Nearest Neighbor. Scholarpedia 4(2):1883. Downloaded 27-04-2014,@ www.google.com.*

27. *Luca, C., Paolo, G.,2013. Improving classification models with taxonomy information. Journal of Data and Knowledge engineering 86(2013) 85-101.[doi:10.1016/j. datak.2013.01.005].*

28. *Michal, M., Jozef, K., Peter, S., 2012. Data preprocessing Evaluation for web log mining: Reconstruction of activities of a web visitor. Journal of procedia computer science. 1(2012) 2273-2280. [doi:10.1016/j.procs.2010.04.255].*

29. *Zdravko, M., Daniel, T. L., 2007. Data mining the Web, Uncovering patterns in Web content, structure, and usage. John Wiley & sons, Inc., New Jersey,USA. P. 115-132.*

30. *Adeniyi, D.A., Wei, Z., Yongquan, Y., 2014. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Journal of Applied Computing and Informatics.. [DOI: 10.1016/j.aci.2014.10.001]*